A Study on the Several Feature Extraction Methods of Protein Classification Using Data Mining Techniques

Purba Mukhopadhyay
Department of Computer Science and Engineering, Brainware University, Barasat, Kolkata, India
2001p.mukhopadhyay@gmail.com

Suprativ Saha[*]
Department of Computer Science and Engineering, Brainware University, Barasat, Kolkata, India
reach2suprativ@yahoo.co.in

Tanmay Bhattacharya
Department of Information Technology, Techno Main, Saltlake, Kolkata, India
dr.tb1024@gmail.com

*Corresponding Author

Abstract:

Protein arrangement is an indispensable field of exploration in Biological Data Mining. The choice of proper elements with a highlight extraction system is a significant piece of this space. These particular elements are applied in any delicate processing strategy to develop a grouping model. In this pa,per several feature extraction procedures like n-gram encoding method, 6-letter exchange group method, frequency based encoding method, extraction based on hydropathy properties, di-sulphide bond, positional average molecular weight, and positional average iso-electric point are described with the proper example. Those feature extraction procedures can produce various feature values from protein which can be applied to any data mining approaches to classify unknown protein in high classification accuracy with low computational time. Furthermore, this paper also provides the various way to classify protein using data mining based on those feature extraction procedure.

Keywords: Protein Classification, Feature Extraction Method, N-gram Encoding Method, Disulphide Bonds, Positional Average Value, Iso-Electric Point Value, Hydropathy Property, 6-letter Exchange Group Method.

Introduction

Data mining is a technique of extracting and discovering patterns from a large amount of data set, and when it is being applied on biological datasets, it is known as biological data mining. In the sphere of biological data mining, the importance and usability of protein classification is spectacular. Protein classification is an approach to classify unknown protein into its class using the sequential and structural properties of the protein. As new protein structures are rapidly increasing day by day, the need for efficient and automated data mining techniques for classifying proteins into classes with high accuracy and low computational time is also increasing. In this scenario, various researches proposed several data mining approaches to classify unknown protein using data mining techniques which is describe in section 2. Section 3 summarized various feature extraction procedure which is used to extract feature values from protein. Finally, a proper conclusion alone with future scope in this area is described in section 4.

Literature Review

In paper [1], *Jason. T. L et. al.* proposed a neural network-based classifier to classify unknown proteins by extracting features using the 2-gram encoding method and 6-letter exchange group method with 90% to 92% accuracy. After that, *Saha. S. et. al.* [2] proposed the saturation point of the n-gram encoding method to reduce the computational time of classification and keep remain the accuracy level. To overcome the various drawbacks of neural-network based classifier, *Mohamed. S. et al.* [3] implemented a fuzzy rule-based model using molecular weight, isoelectric point, hydropathy properties as the features for classifying unknown protein with 93% high accuracy. In [4] *Saha. S. et al.* applied the positional weighted average molecular weight and positional average iso-electric value to the Fuzzy ARTMAP model to increase the accuracy. To handle the large amount of data and to identify the necessary features,a rough set approach was provided a new classifier by Pawlak [5].

Hybridization approach may be solved the various problem based on accuracy and computational time of protein classification, which are generated by a previous non-hybrid classifier. In paper [6] *Sen S. et al.* proposed a python based standalone tool i.e., PyPredT6 to predict the T6 effector proteins. After that *Saha S. et al.* [7][8] proposed a feature grouping hybridization procedure that involve 3 phases like the combination of neural network system, fuzzy ARTMAP model and rough set classifier. In this procedure, *KMP algorithm* was applied to reduce the computational time with an accuracy 91%. Frequency based encoding method was proposed by *Iqbal. M. J. et al.* [9] for classifying the protein sequence and determining their structure and function. Beside the behavioral approaches of classification, structural feature extraction procedure can provide valuable techniques in protein classification. In this case, *Iqbal. M. J. et al.* [10] proposed a distance-based encoding method with 91.2% accuracy where the features are extracted from the input protein sequences and find the distance between the occurrences of the same amino acid which is used as a feature value and tested with different classifiers.

Jiang Qiangrong et al. [11] used a graph kernel-based model combining with the neural network on protein classification. During the time of drug design, the important task is to study and classify the unknown protein into a known protein family. *Babasaheb S. Satpute et al.* [12] proposed a probabilistic approach involving feed-forward and feedback ANN, Neive based, SVM, and Decision tree to classify protein with efficiency of 63%, 59%, 68%, and 84% respectively. During the case of bacterial identification and bacterial protein detection MALDI-TOF is a rapid sensitive technique. *Tomachewski D. et al.* [13] developed a tool i.e., Ribopeaks, for bacterial classification through m/z data from ribosomal protein with the database of more than 28,500 bacterial taxonomic records. To classify circular RNA from other long non-coding RNA, Benson, D.A. et al. [15] and Chaabane M. et al. [14] proposed the Reverse Complement Matching (RCM) descriptor and ACNN-BLSTM sequence descriptor combines the asymmetric convolution neural network (ACNN) with the Bidirectional Long Short-Term Memory network (BLSTM) where the shared representations across different modalities are integrated.

Identification of protein similarity is an important task for protein sequence classification and homology detection. *Spalding J.D. et al.* [16] proposed the string kernel method-based classifier which developed a strategy for efficient estimation of suitable kernel parameter values. Here the Kullback-Leibeir (KL) distance was calculated between the observed k-mar frequencies and the theoretical k-mar frequencies of protein data. *A.F. Ali et al.* [17] predicted the functional classification of protein sequences based on a set of features involving Fast Fourier transformation (FFT) of molecular weight of each protein sequence were applied on SCOP database. *Cornelia Caragea et al.* [18] proposed feature hashing technique, to reduce the complexity of learning algorithms where input belongs to high dimensional space. *Robert Busa-Fekete et al.* [19] proposed the phylogenetic analysis approach with 93% accuracy, followed by Tree Insert and TreNN algorithm for protein classification.

Pranay Desai [20] had proposed Hidden Markov Models based classifiers with 94% accuracy, which were performed in three phases such as training, decoding, and evaluation to identify functional properties of input data. Selecting the most informative features and reducing the dimensionality of the feature vector is an important task in protein sequence classification. *Xing-Ming Zhao et al.* [21] proposed a classifier with the combination of Genetic algorithm and support vector machine framework. Protein structural classification is another important classification approach to classify protein based on protein chemical structure. *Rahman M. M. et al.* [22] had proposed hierarchy tree structure with six major features of a protein like Sequence Comparison, Structure comparison, Cluster Index, Connectivity, Taxonomic and Interactivity with 98% accuracy.

Feature Extraction Procedure

To classify the protein sequence, features must be extracted from the input data. So, here comes the need of feature extraction method. Researchers have done many research works in this field. There are various popular feature extraction methods like n-Gram encoding method, di-sulphide bonds etc. Here, in this paper, the discussion over some popular feature extraction methods has been done.

N-Gram Encoding Method: A Feature Extraction Approach

A protein sequence contains the combination of twenty amino acids which is recognized by twenty letters of English alphabets. The N-gram encoding method is highly appreciable approach, which is used in neural network-based classifier to extract features from the protein sequences. In N-gram encoding method, value of 'N' can be varied from 2 to n. Individual features are extracted in every gram value. At first, occurrence of amino acid is calculated where window size is 'n'. After that, mean value and the standard deviation are generated based on the occurrence of amino acid group using the following formulas [eq. 1 & eq. 2] where 'MN' denotes mean value and 'SD' denotes standard deviation.

$$\text{MN} = \frac{\sum_{k=1}^{f} \quad z_k}{f} \qquad (1)$$

$$SD = \sqrt{\frac{(\sum_{k=1}^{f} \quad (z_k - MN)^2)}{(f-1)}} \qquad (2)$$

Now, in this case, it is important to maintain high accuracy level and low computational time for efficiency purpose and for obtaining optimum result. So, to maintain this, it needs to fix the upper limit of 'n'. Also, standard deviation is one of the most important features, which can be obtained by two different methods (using standard mean value and floating mean value). Now, from previous research works done by *Saha et al.* [2], it can be concluded that calculation of standard deviation using floating mean value is more significant than calculation using standard mean value. On the other hand, it is noticed that, after 5-gram, from 6-gram to n-gram all the value of standard deviation in both procedures is bounded to zero. So, in this case, it can be also concluded that the saturation point of the N-gram encoding method is fixed to 5-gram. Thus, a significant improvement in the time of execution without hampering the accuracy level of classification is obtained.

Feature Extraction by using di-Sulphide bonds

In the field of bioinformatics, feature extraction using di-sulphide bonds is another most effective method. Protein disulphide bonds are the links between pairs of cysteine residues in the polypeptide chain. These bonds are classified based on the sign of the five dihedral angles that define the cystine residue. Twenty disulphide conformations are possible using this convention and all 20 are represented in protein structures. However, many research works have been done in this field and several pre-existing classifiers recognized the use of a single type of Disulphide bond (viz, parallel, or alternate) as a useful feature. In this basis, experiments about this and various combinations of disulphide bonds had been studied to formulate a potent protein feature, after that a data mining approach had been applied on the seven different combinations of disulphide bonds (viz. parallel [eq. 3], alternate [eq. 4] and quad [eq. 5]) to identify the best feature. After the experiment, it can be seen that with respect to all the other combinations of disulphide bonds, the combination of alternate–quad bonds turned out to be the best and most efficient feature and its accuracy level of classification had extended high up to 93%. So, it is revealed that the combination of alternative and quad disulphide bonds can be used as an effective feature in any form of protein classification.

$$d_p = \sum_{i=1}^{n} \frac{P_{i+1} - P_i}{count} \qquad (3)$$

$$d_a = \sum_{i=1}^{n} \frac{(P_{i+2} - P_i) + (P_{i+3} - P_{i+1})}{\frac{count}{2}} \qquad (4)$$

$$d_q = \sum_{i=1}^{n} \frac{(P_{i+3} - P_i) + (P_{i+2} - P_{i+1})}{\frac{count}{2}} \qquad (5)$$

Feature Extraction Using Positional Average Value

To classify unknown protein sequences into proper class, subclass and family, different features are extracted from the protein sequences, which are applied on any popular soft computing methodology. The most popular features used for protein sequence classification are average molecular weight and iso-electric point value, which are applied to Fuzzy ARTMAP model. But some weakness which may affect the efficiency and accuracy of this model is found in these two approaches. To overcome this, in the research work done by *Saha.et.al.* [4], four groups of feature extraction procedures with the combination of positional and non-positional average values of features was proposed which is applied in fuzzy ARTMAP model individually to classify unknown protein to its family. They worked with 497 unknown sequences of six different families to identify the best group among all in the basis of classification accuracy as well as computational time. As a result, it is noticed that, to avoid the weakness of Fuzzy ARTMAP model, position value should be multiplied with the value of molecular weight or iso-electric point of every individual amino acid in a protein sequence, and also the combination of positional-average molecular weight [eq. 6] and positional-average isoelectric point [eq. 7] has provided the most significant result of classification to increase the accuracy level and efficiency of classification.

$$PAMW = \frac{\sum_{i=1}^{n} m_i * i}{l} \qquad (6)$$

$$PAIE = \frac{\sum_{i=1}^{n} s_i * i}{l} \qquad (7)$$

Feature extraction using Hydropathy Property of protein

Among the sequence properties, the hydropathy distribution in proteins (the patterns of hydrophilicity and hydrophobicity) has been used extensively for protein structure prediction and structural classification of proteins. In previous research works, calculation is being done only using two methods i.e. hydropathy composition(C) and hydropathy transmission (T).

Hydropathy Composition

Hydropathy composition (the frequency of each 20 possible amino acids) can present in a sequence, hydrophobic, hydrophilic and neutral. So, using the below formulas, frequency of hydrophobic [eq. 8], hydrophilic [eq. 9] and neutral amino acids [eq. 10] can be calculated.

$$PHPHO = \frac{\sum_i^l \quad hpho_i * 100}{l} \tag{8}$$

$$PHPHI = \frac{\sum_i^l \quad hphi_i * 100}{l} \tag{9}$$

$$PNEU = \frac{\sum_i^l \quad neu_i * 100}{l} \tag{10}$$

Hydropathy Transmission

Hydropathy transmission can be defined by three values, first, the number of occurrences of hydrophilic molecule followed by neutral molecule and vice versa, second, neutral molecule is followed by a hydrophobic molecule or vice versa and third, the hydrophilic molecule is followed by a hydrophobic molecule or vice versa. Also, another three combinations can be calculated here, where a hydrophobic molecule is followed by a hydrophobic molecule, a hydrophilic molecule is followed by a hydrophilic molecule and a neutral molecule is by a neutral molecule.

6-Letter Exchange Group Method

6-letter exchange group method is utilized to represent a protein sequence. It plays a vital role in extracting features from unknown protein sequences. Here, 6-letter exchange group {e1, e2, e3, e4, e5, e6} is adopted to represent a protein sequence, where e1€ {H,R,K}, e2€{D,E,N,Q}, e3€{C}, e4€{S,T,P,A,G}, e5€{M,I,L,V} and e6€{F,Y,W}. For example, the protein sequence MALRKECT can be represented as e5e4e5e1e1e2e3e4. Initially, 2-gram encoding exchange group method is applied on the converted sequence. After that, the mean value, the standard deviation and the coefficient of variance are generated based on the occurrence of two consecutive exchange group patterns using formulas (Eq. 12) used for n-gram encoding method. Later, the calculated mean value, standard deviation value and coefficient of variance of occurrence are normalized by applying the following formulas.

$$Sig_x = \frac{1}{1 + e^{-x}} \tag{12}$$

Where Sig_x represent the sigmoid or normalised value

Frequency Based Encoding Method

Frequency Encoding is an encoding technique which encodes categorical feature values to their frequencies. It determines the occurrence probability of each amino acid in a sequence. From the following formula, occurrence probability can be calculated.

$$F_{ix} = \frac{N(x|S_j)}{l} \qquad (13)$$

Where, $N(x|S_j)$ is the total number of occurrence of each amino acid, l is the length

For example, there is a sample protein sequence MALCAKML of length 8. So, the result is

Amino Acids	M	A	L	C	K
Frequency Value	2	2	2	1	1
Occurrence Probability	0.25	0.25	0.25	0.125	0.125

Tab1. Result of fequency based encoding method

Conclusion

In this paper, we have discussed about several feature extraction methods that are used to classify protein sequences. The most important common part and the ultimate moto of all these feature extraction methods is to classify unknown protein sequence with high accuracy level and low computational time, and to find out an optimal result. But, it will be more better if this can happen with less number of methods. If we can decrease the the number of features i.e. dimensions , then, the computational time will be less than before and the work will be easier. Many researchers have worked upon this field and day by day it's improving a lot. In future, many experiments will be done using dimension reduction method over these feature extraction methods for more efficiency purpose.

References

1. Wang, J. T. L., Ma, Q., Shasha, D., & Wu, C. H. (2000). Application of neural networks to biological data mining: A case study in protein sequence classification. In R. Ramakrishnan, S. Stolfo, R. Bayardo, I. Parsa, R. Ramakrishnan, S. Stolfo, R. Bayardo, & I. Parsa (Eds.), *Proceeding of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 305-309).
2. Saha. S and Bhattacharya. T (2018) An Approach to Enhance the Design of Protein Sequence Classifier Using Data Mining, *Procedia Computer Science 167* (2020) 717–726.
3. Shakir Mohamed, David Rubin and Tshilidzi Marwala (2006) Multi-class Protein Sequence Classification Using Fuzzy ARTMAP. *IEEE Conference*,1676 – 1680.
4. Saha. S and Bhattacharya. (2018) T A new Protein Sequence Classification approach using Positional Average Values of features, *AISC, Springer, SoCTA2018.*
5. Z. Pawlak (2002) Rough set theory and its applications, J. Telecommun. Inf. Technol
6. Rishika Sen, Losiana Nayak and Rajat K. De (2019), A python-based prediction tool for identification of Type VI effector proteins JBCB, vol:17,PP: 1950019-1 to 1950019-17
7. Suprativ Saha and Rituparna Chaki (2013) "A Brief Review of Data Mining Application Involving Protein Sequence Classification", AISC, Springer, Volume 177, ACITY 2012, Chennai, India, pp. 469-477.

8. Suprativ Saha and Rituparna Chaki (2012) “Application of Data Mining in Protein Sequence Classification”, IJDMS, Volume 4, Number 5, October 2012, AIRCC, DOI: 10.5121/ijdms.2012.4508, pp. 103-118, ISSN: 0975-5705 (Online), 0975-5985 (Print)
9. Muhammad Javed Iqbal, Ibrahima Faye, Abas Md Said, Brahim Belhaouari Samir (2014) “An Efficient Computational Intelligence Technique for Classification of Protein Sequences”,IEEE 2014, pp. 1-6
10. Muhammad Javed Iqbal, Ibrahima Faye, Abas Md Said and Brahim Belhaouari Samir (2013) “A Distance-Based Feature-Encoding Technique for Protein Sequence Classification in Bioinformatics”, CYBERNETICSCOM 2013, IEEE 2013, pp. 1-5
11. Jiang Qiangrong and Qiu Guang (2019), Graph kernels combined with the neural network on protein classification, JBCB, vol.17, PP:1950030-1-1950030-11
12. Babasaheb S. Satpute and Raghav Yadav (2019), An Efficient Machine Learning Technique for Protein Classification Using Probabilistic Approach, AISC, Springer, Vol. 828, PP: 405 – 413
13. Tomachewski D.et al (2018), Ribopeaks: a web tool for bacterial classification through m/z data from ribosomal proteins, Bioinformatics, 34(17), 2018, 3058–3060.
14. Chaabane, M. et al (2018), circDeep: deep learning approach for circular RNA classification from other long non-coding RNA, Bioinformatics, 36(1), 2020, 73–80.
15. Benson, D.A. et al. (2017) GenBank. Nucleic Acids Res., 45, D37.
16. Spalding J.D. and Hoyle D.C. (2005) Accuracy of String Kernels for Protein Sequence Classification, ICAPR 2005. Springer (LNCS) vol 3686.
17. A.F. Ali and D. M. Shawky (2010) A Novel Approach for Protein Classification Using Fourier Transform, IJEAS, 6:4 2010.
18. Cornelia Caragea, Adrian Silvescu and Prasenjit Mitra (2012) Protein Sequence Classification Using Feature Hashing, Proteome Sci. 2012; 10 (Supple 1): S14.
19. Robert Busa-Fekete, Andras Kocsor, and Sandor Pongor (2010) Tree-Based Algorithms for Protein Classification. International Journal on Computer Science and Engineering (IJCSE).
20. Pranay Desai (2005) Sequence Classification Using Hidden Markov Model.
21. Xing-Ming Zhao et al. (2004) A Novel Hybrid GA/SVM System for Protein Sequences Classification, IDEAL 2004, Springer (LNCS) 3177, pp. 11-16.
22. Muhammad Mahbubur Rahman, Arif Ul Alam, Abdullah-Al-Mamun and Tamnun E Mursalin (2011) A More Appropriate Protein Classification Using Data Mining, (JATIT), pp. 33-43